CLAIMS

What is claimed is:

1. A method for recognizing the structure of a delineated table region in an electronic document, comprising the steps of:

a) creating a binary tree using a hierarchical clustering of a plurality of words included in said table region;

b) segregating a plurality of table columns using a breadth-first traversal algorithm;

c) identifying column headers, if any, using a first heuristic algorithm; and

d) identifying row headers, if any, using a second heuristic algorithm; and

e) segregating at least one table row using a row determination algorithm.

2. The method according to claim 1, wherein the hierarchical clustering further comprises the steps of:

a) generating a plurality of leaf clusters;

b) calculating a plurality of inter-cluster distances for each one of a plurality of clusters;

b) merging the two clusters having a minimum inter-cluster distance calculated in b) to create a new cluster;

c) creating an interior node of the binary tree with the said two clusters as its children; and

d) repeating steps b) through d) until there is only one cluster left without a parent.

3. The method according to claim 2, wherein each one of the plurality of leaf clusters comprises a single one of said plurality of words

4. The method according to claim 2, wherein the inter-cluster distance is determined by an algorithm comprising the steps of:

a) calculating a position vector (span) for each one of the plurality of words, said span comprising the starting and ending horizontal position of each said word; and

b) determining a unique separation distance between each unique cluster of the plurality of clusters and each one of the other clusters in the plurality of clusters by:

1) using positional vector subtraction of the individual cluster positional vectors when each cluster is comprised of a single word; and

2) when at least one of the clusters is a merged cluster, computing the average separation distance of all the unique inter-cluster separation distances comprising the cluster pair.

5. The method according to claim 4, wherein the distance comprises one from the group consisting of geometric, syntactic, and semantic.

6. The method according to claim 1, wherein the breadth-first traversal algorithm comprises the steps of:

a) beginning at the root node, split the node into two nodes and determine whether the two split nodes can be split into subordinate nodes based on a spacing decision criteria;

b) if a node cannot be split, move the node into a storage buffer, else repeat step a for any remaining nodes; and

c) when all nodes have been moved into the storage buffer, the columns are defined as the nodes in the storage buffer.

7. The method according to claim 6, wherein the spacing decision criteria comprises the splitting of the node if and only if:

a) the node is the root node;

b) if $g \geq G$; or

c) if $g < G$ and $g/m_g > \alpha$

where g is a gap between clusters, G is a predetermined constant, m_g is an average gap between adjacent pairs of already identified columns, and α is a number between 0 and 1.

8. The method according to claim 6, additionally including the step of sorting columns according to a starting position of each one of the plurality of columns.

9. The method according to claim 8, additionally including the step of adjusting the upper boundary of the table region by performing a consistency test.

10. The method according to claim 9, wherein the consistency test comprises the steps of:

a) calculating a predominate string type for each one of the plurality of columns included in the table region (column type);

b) starting at a predetermined number of table lines below the start of the table region, calculating a unique string type for each one of the plurality of words in said table line (word type);

c) comparing each one of the plurality of word types with the associated column type;

d) generating a plurality of metrics associated with the result of said comparisons; and

e) if a majority of said metrics are true, identifying the current line as the bottom line of the box region and ending the consistency test, or else moving up one table line and repeating steps c) and d).

11. The method according to claim 1, wherein the first heuristic algorithm for identifying column headers comprises the steps of:

a) dividing each table line into a plurality of unique separable strings;

b) creating a hierarchical tree having a box as the root, each one of the plurality of table columns as the leaves, and higher level headers as intermediate nodes of said tree;

c) calculating a joint span for each one of the plurality of separable strings using the equation

$$p_{1,n} = (\min(s_i), \max(e_i)) \qquad i = 1 \text{ to } n$$

d) comparing the boundaries of each one of the plurality of joint spans with the boundaries of each one of the plurality of table columns; and

e) creating a list of associated columns that have overlapping boundaries in b) using a boundary criteria.

12. The method according to claim 11, wherein each one of the separable strings are delineated by a predetermined number of blank spaces.

13. The method according to claim 11, wherein the boundary criteria further comprises the steps of:

a) associating each phrase with at least one column; and

b) if a phrase is associated with more than one column, the subsidiary columns must already have its own header filled.

14. The method according to claim 1, wherein the second heuristic algorithm for identifying row headers comprises the steps of:

a) identifying a region as a stub region if the left-most column does not include a column header;

b) performing a semantic analysis of the data contents of the left-most column if the left-most column does include a column header; and

c) detecting and storing the unique row headers for each line from steps a) and b).

15. The method according to claim 1, wherein the row determination algorithm comprises the steps of:

a) defining a row separator if said row comprises a blank line;

b) determining at least one core row, comprising:

1) a row having a non-empty string in a stub region and having at least one other column; or

2) a row having non-empty strings in a majority of the columns of the table; and

c) determining non-core rows, if any.

16. The method according to claim 15, wherein non-core rows comprise all rows that are not core rows.

17. The method according to claim 1, additionally including the step of testing of said delineated table by creating a directed acyclic graph.

18. The method according to claim 17, additionally including the step of testing of said delineated table logically probing said directed acyclic graph.

19. The method according to claim 18, wherein the step of testing said table comprises the comparison of responses from a plurality of logical tests conducted on said graph with an associated plurality of predetermined reference responses.

20. A method for querying an electronic table comprising the steps of:

a) creating and storing a first list of keywords, said keywords representing the Acells of a table;

b) creating and storing a second list of keywords to be used to determine actions to be taken with said table;

c) parsing said query for at least one action keywords that matches at least one word included in said second list of keywords;

d) parsing said query for at least one keyword that matches at least one word in the first list of keywords.

21. The method according to claim 20, wherein said first list of keywords comprises the plurality of words included in a box and a stub regions of said table.